On density-based data streams clustering algorithms: A survey.

Teh Wing Wah*
Faculty of Computer Science an Information Technology, University of Malaya
Email: tehyw@um.edu.my

Abstract: Clustering data streams has drawn lots of attention in the few years due to their ever-growing presence. Data streams put additional challenges on clustering such as limited time and memory and one pass clustering. Furthermore, discovering clusters with arbitrary shapes is very important in data stream applications. Data streams are infinite and evolving over time, and we do not have any knowledge about the number of clusters. In a data stream environment due to various factors, some noise appears occasionally. Density-based method is a remarkable class in clustering data streams, which has the ability to discover arbitrary shape clusters and to detect noise. Furthermore, it does not need the number of clusters in advance. Due to data streams characteristics, the traditional density-based clustering is not applicable. Recently, a lot of density-based clustering algorithms are extended for data streams. The main idea in these algorithms is using density-based methods in the clustering process and at the same time overcoming the constraints, which are put out by data stream's nature. The purpose of this paper is to shed light on some algorithms in the literature on density-based clustering over data streams. We not only summarize the main density-based clustering algorithms on data streams, discuss their uniqueness and limitations, but also explain how they address the challenges in clustering data streams. Moreover, we investigate the evaluation metrics used in validating cluster quality and measuring algorithms' performance. It is hoped that this survey will serve as a steppingstone for researchers studying data streams clustering, particularly density-based algorithms.

Keywords: Data stream; density-based clustering; grid-based clustering; micro-clustering.